Open Research Online

The Open University's repository of research publications
and other research outputs

Modeling Population Structure Under Hierarchical Dirichlet Processes

Journal Item

How to cite:

Elliott, Lloyd T.; De Iorio, Maria; Favaro, Stefano; Adhikari, Kaustubh and Teh, Yee Whye (2019). Modeling Population Structure Under Hierarchical Dirichlet Processes. Bayesian Analysis, 14(2) pp. 313–339.

For guidance on citations see FAQs.

© 2019 International Society for Bayesian Analysis

https://creativecommons.org/licenses/by/4.0/

Version: Version of Record

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1214/17-BA1093

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

Bayesian Analysis (2019) **14**, Number 2, pp. 313–339

Modeling Population Structure Under Hierarchical Dirichlet Processes

Lloyd T. Elliott[*], Maria De Iorio[†], Stefano Favaro[‡,§], Kaustubh Adhikari[¶], and Yee Whye Teh[‖,**]

Abstract. We propose a Bayesian nonparametric model to infer population admixture, extending the hierarchical Dirichlet process to allow for correlation between loci due to linkage disequilibrium. Given multilocus genotype data from a sample of individuals, the proposed model allows inferring and classifying individuals as unadmixed or admixed, inferring the number of subpopulations ancestral to an admixed population and the population of origin of chromosomal regions. Our model does not assume any specific mutation process, and can be applied to most of the commonly used genetic markers. We present a Markov chain Monte Carlo (MCMC) algorithm to perform posterior inference from the model and we discuss some methods to summarize the MCMC output for the analysis of population admixture. Finally, we demonstrate the performance of the proposed model in a real application, using genetic data from the ectodysplasin-A receptor (EDAR) gene, which is considered to be ancestry-informative due to well-known variations in allele frequency as well as phenotypic effects across ancestry. The structure analysis of this dataset leads to the identification of a rare haplotype in Europeans. We also conduct a simulated experiment and show that our algorithm outperforms parametric methods.

Keywords: admixture modeling, Bayesian nonparametrics, hierarchical Dirichlet process, linkage disequilibrium, population stratification, single nucleotide polymorphism data, MCMC algorithm.

1 Introduction

Population stratification, also known as population structure, refers to the presence of a systematic difference in genetic markers' allele frequencies between subpopulations due to variation in ancestry. This phenomenon arises from the bio-geographical distribution of human populations. The analysis of population structure represents an important problem in population genetics. Broadly speaking, this analysis can solve problems related to: *i)* detecting population structure; *ii)* estimating the number of subpopulations in a sample; *iii)* assigning individuals to subpopulations; *iv)* defining the number of

[*]Department of Statistics, University of Oxford, Oxford OX1 3TG, U.K., elliott@stats.ox.ac.uk
[†]Department of Statistical Science, University College London, London WC1E 6BT, U.K., m.deiorio@ucl.ac.uk
[‡]Department of Economics and Statistics, 10134 Torino, Italy
[§]Collegio Carlo Alberto, 10024 Moncalieri, Italy, stefano.favaro@unito.it
[¶]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, U.K., k.adhikari@ucl.ac.uk
[‖]Department of Statistics, University of Oxford, Oxford OX1 3TG, U.K.
[**]Google DeepMind, London EC4A 3TW, U.K., y.w.teh@stats.ox.ac.uk

© 2019 International Society for Bayesian Analysis https://doi.org/10.1214/17-BA1093

ancestral populations in admixed populations; *v)* inferring ancestral population proportions to admixed individuals; *vi)* identifying the genetic ancestry of distinct chromosomal segments within an individual. These analyses are crucial to the understanding of human migratory history and the genesis of modern populations (*e.g.*, Rosenberg et al. (2002) and Reich et al. (2009)). In particular the associated admixture analysis of individuals is important in correcting the confounding effects of population ancestry on gene mapping (Zhu et al., 2008) and association studies (Price et al., 2010). It is also useful in the analysis of gene flow in hybridization zones (Field et al., 2011) and invasive species (Ray and Quader, 2014), conservation genetics (Wasser et al., 2007) and domestication events (Parker et al., 2004). The establishment of inexpensive single nucleotide polymorphism (SNP) genotyping platforms in recent years has facilitated the collection of markers to assess genetic ancestry in human populations and in general to investigate genetic relationships in living organisms. This paper focuses on a particular form of population structure, which is known as admixture. Genetic admixtures occur when two or more previously isolated populations begin interbreeding, resulting in the introduction of new genetic lineages into a population (*e.g.* the African-American population).

A variety of modeling approaches have been proposed for the analysis of population structure. Two of the most widely used approaches are principal component analysis (PCA) and model-based estimation of ancestry, mainly involving clustering techniques or hidden Markov models (HMM). The PCA approach has been used to infer population structure for several decades. In the PCA approach, the individuals' genotypes are projected onto a lower dimensional space so that the locations of individuals in the projected space reflects their genetic similarities (*e.g.*, Patterson et al. (2006) and Novembre and Stephens (2008)). It should be noted that the top principal components do not always capture population structure but may reflect family relatedness, long range linkage disequilibrium or simply genotyping artefacts. Model-based estimation methods aim to reconstruct historical events and therefore to infer explicit genetic ancestry (*e.g.*, Pritchard et al. (2000), Tang et al. (2005) and Alexander et al. (2009)). In the structured association approach samples are assigned to subpopulation clusters, possibly allowing for fractional membership. Among model-based estimation methods, an influential early approach is STRUCTURE. This approach was proposed by Pritchard et al. (2000) and it assumes that individuals come from one of K subpopulations. Based on Bayesian mixture models, population membership and population specific allele frequencies are jointly estimated from the data. This simple framework can be extended to genetic admixtures, allowing individuals to have ancestry from more than one population. For each individual, STRUCTURE estimates what proportion of the individual's genome comes from each population, while the alleles at different loci are modeled as conditionally independent given these admixture proportions. Taking a Bayesian approach to inference, independent priors on the allelic profile parameters of each population are specified and posterior inference is performed through MCMC.

One problem with STRUCTURE, which we address in this paper, is that of admixture linkage disequilibrium among neighbouring loci. When individuals from different groups admix, their offsprings' DNA become a mixture of the DNA from each admixing group. Chunks of DNA are passed along through subsequent generations, up to the time of sample collection. Therefore, the genomes of the descendants contain segments

of DNA inherited from each of the original populations. The shorter the distance between two loci, the higher the probability that the population of ancestry will be the same at these two loci. This means that ancestry states are autocorrelated. The lengths of uninterrupted DNA segments inherited from each subpopulation reflect how long ago the admixture event occurred. In general long uninterrupted segments from each population imply a recent admixture event. The original version of STRUCTURE did not deal with admixture linkage disequilibrium and as a result it is necessary to thin out tightly-linked loci to reduce correlations which can affect the quality of inference. Falush et al. (2003) improved on this issue (with STRUCTURE2) by introducing a module to model linkage locally among neighbouring loci, using a Markov model which segments each chromosome into contiguous regions with shared genetic ancestry. This allows local genetic ancestry from genotype data to be inferred, as opposed to the global admixture proportions in Pritchard et al. (2000). Such a local ancestry estimation then gives more fine-grained information about the admixture process.

Another important statistical concern in admixture modeling, which we address in this paper, is the determination of the number of ancestral populations. In Pritchard et al. (2000), this is achieved using a model selection criteria based on MCMC estimates of the log marginal probabilities of the data and the Bayesian deviance information criterion, though it has been noted by Falush et al. (2003) that such estimates are highly sensitive to prior specifications regarding the relatedness of the populations. See also Corander et al. (2003) and Evanno et al. (2005) for other parametric approaches to determining the number of populations. One way in which such model selection can be sidestepped is by using a Bayesian nonparametric approach, which offers a flexible framework and does not require the specification of a fixed and finite model size. Rather, one assumes an unbounded potential model size, of which only a finite part is observed on a given finite dataset. See the monograph by Hjort et al. (2010) and references therein for a comprehensive and stimulating account of the Bayesian nonparametric approach. In the population structure setting, the model size is the number of the populations. In the analysis of population structure, the idea of assuming an unbounded potential number of populations was first considered by Huelsenbeck and Andolfatto (2007). They used a Dirichlet Process (DP) to define a Bayesian nonparametric counterpart to the "no-admixture" model of Pritchard et al. (2000). See also Dawson and Belkhir (2001) and Pella and Masuda (2006) for extensions to polyploid data.

In this paper we introduce a new method for modeling population structure that simultaneously gives estimates of local ancestries and bypasses difficult model selection issues using a Bayesian nonparametric approach. In other terms the proposed model provides a Bayesian nonparametric counterpart of the admixture model by Falush et al. (2003). Our approach relies on the hierarchical Dirichlet process (HDP) by Teh et al. (2012), which models the unknown and uncertain number of populations without having to perform costly model selection. This is combined with a transition model in which, during a transition event, the founder identities on either side of the transition are independent. This transition model requires a linear (in the number of founders) number of parameters, as well as a forward-backward algorithm which scales linearly in the number of extant populations while introducing fewer auxiliary variables which can slow

down convergence. We illustrate our approach on a sample of 372 Colombian people, which have a well-known history of genetic admixture.

The history of Bayesian nonparametric models for genetic variation began with work in which a general HDP-HMM (the infinite HMM) was used to model haplotypes (Sohn and Xing, 2007). Since then, Bayesian nonparametric versions of the popular fastPHASE and BEAGLE models have been developed (Elliott and Teh, 2012, 2016), as well as extensions of Sohn and Xing (2007) for known population structures. It is well known that clustering methods based on Dirichlet processes have the same statistics as clusterings induced by Kingmans' coalescent (Neal, 2003), and so Bayesian nonparametrics provide a natural framework for generalising parametric HMMs for haplotype models. Furthermore, Bayesian nonparametric approaches allow some of the ad-hoc aspects of parametric models (such as the choice of the number of latent populations used in an admixture) to be replaced with theory.

Our model extends the above research by providing a Bayesian nonparametric version of STRUCTURE2. The main challenge in the specification of this model is the development of inference, which we have done with a novel MCMC method which allows efficient exact inference (up to MCMC sampling error) using a retrospective slice sampling truncation scheme. In order to capture the nonhomogeneous nature of allele emission probabilities, standard MCMC for Bayesian nonparametrics cannot be used. Consequently, our inference methods are also relevant for other work in Bayesian nonparametrics HMMs in which nonhomogeneous behaviour is required. For example, the novel MCMC techniques developed in this paper could also be used to extend Beta Process HMMs for video capture (Fox et al., 2012) for the use with time varying contexts or covariates (*i.e.*, sections of the video capture could be done in different and known lighting conditions).

In Section 2 we introduce our Bayesian nonparametric model and inference method. Section 3 describes population structure analyses of genotype data from the EDAR gene region, and also a simulated experiment in which we compare our analysis and algorithm runtimes to STRUCTURE2 (Falush et al., 2003) and also a baseline. Section 4 closes with a discussion of our findings as well as potential future work.

2 Method

We assume multilocus genotype data from a sample of admixed individuals arising from a number of populations. For simplicity, we assume N haploid individuals genotyped at L loci, and we denote by $X = (x_{il})_{1\leq i\leq N, 1\leq l\leq L}$ the observed data, where x_{il} is the allele of individual i at locus l.

2.1 Model and prior specification

Let K be the number of ancestral populations. We denote by $Q_i = (q_{ik})_{1\leq k\leq K}$ the vector of admixture proportions of individual i, where q_{ik} denotes the proportion of the genome of individual i which can be traced to population k. While previous works used

a finite value for K, we will take a Bayesian nonparametric approach and let $K \to \infty$, so that there is an unbounded number of potential populations in the model. To account for dependence among loci, we use the model of linkage proposed by Falush et al. (2003). This employs a hidden Markov model which splits the genome into contiguous chunks with common ancestry. The model is parameterised by: d_l, the genetic distance between locus l and locus $l+1$, for each $l = 1, \ldots, L-1$, and r, the rate at which splits occur. Let z_{il} be a variable which denotes the population ancestry at locus l of individual i, and s_{il} be a binary variable which denotes whether locus $l-1$ and locus l are in the same chunk ($s_{il} = 1$) or not ($s_{il} = 0$). We define $s_{i1} = 0$ for all i. The variables s_{il} can be thought of as linkage indicator variables. In particular the transition model is defined as follows

$$\begin{aligned} z_{i1} &\sim \text{Discrete}(Q_i), \\ s_{i,l+1} &\sim \text{Bernoulli}(e^{-rd_l}), \quad l = 1, \ldots, L-1, \\ z_{i,l+1}|s_{i,l+1}, z_{il} &\begin{cases} = z_{il} & \text{if } s_{i,l+1} = 1, \\ \sim \text{Discrete}(Q_i) & \text{if } s_{i,l+1} = 0. \end{cases} \end{aligned} \tag{1}$$

The probability of a split between loci l and $l+1$ is $1 - e^{-rd_l}$, and the ancestral populations of each chromosome segment are independent and identically distributed. The probability that the ancestral population of a chunk is the k-th ancestral population is q_{ik}.

The model is completed by specifying the likelihood function for the observed alleles. We will assume that within each population Hardy–Weinberg equilibrium holds, and we can model the allelic profile of the kth population simply by specifying the vector of allele frequencies, that is $\theta_k = (\theta_{kla})_{1 \le l \le L, 1 \le a \le A}$, where θ_{kla} is the probability for allele a at locus l in population k. That is,

$$x_{il}|z_{il} = k \sim \text{Discrete}(\theta_{kl}), \tag{2}$$

where $\theta_{kl} = (\theta_{kla})_{1 \le a \le A}$. For single nucleotide polymorphism (SNP) data, x_{il} are binary valued and modeled by Bernoulli distributions with means given by θ_{kl1}. The admixture model of Pritchard et al. (2000) can be recovered from (2) as $r \to \infty$, as all loci become independent and the chunks consist of single loci.

The typical prior in previous works (*e.g.*, Balding and Nichols (1995); Rannala and Mountain (1997); Pritchard et al. (2000) and Falush et al. (2003)) is given by a symmetric Dirichlet distribution, which assumes that all populations have *a priori* equal contribution to the observed genomes. Here we use an asymmetric Dirichlet with mean $Q_0 = (q_{0k})_{1 \le k \le K}$ instead, to capture the assumption that some populations may be more prevalent than others, so have *a priori* higher chances of contributing more genetic material to each individual. See, *e.g.*, Anderson (2001) and Anderson and Thompson (2002). In particular we assume that

$$Q_i|Q_0 \sim \text{Dirichlet}(\alpha Q_0), \tag{3}$$

where $\alpha > 0$ is a parameter which controls the concentration of the Dirichlet prior around Q_0.

The asymmetric Dirichlet also allows for a Bayesian nonparametric model, in which the number of populations K is taken to be infinite, while the corresponding infinite K limit does not lead to a mathematically well-defined model for the symmetric Dirichlet. Specifically, consider a hierarchical prior on Q_0 expressed in terms of the stick-breaking representation by Sethuraman (1994), *i.e.*,

$$\text{For } k = 1, 2, \ldots: \qquad \begin{aligned} v_{0k} &\sim \text{Beta}(1, \alpha_0), \\ q_{0k} &= v_{0k} \prod_{k'=1}^{k-1} (1 - v_{0k'}), \end{aligned} \tag{4}$$

where α_0 controls the overall diversity of populations, with larger α_0 corresponding to a larger number of populations with more uniform proportions. The conditional distribution of Q_i given Q_0 is still a Dirichlet as given in (3), though we need to extend the definition to infinite-dimensional vectors. A constructive definition of such an infinite-dimensional Dirichlet is given as follows

$$\text{For } k = 1, 2, \ldots: \qquad \begin{aligned} v_{ik} &\sim \text{Beta}(\alpha v_{0k}, \alpha(1 - \sum_{k'=1}^{k} v_{0k'})), \\ q_{ik} &= v_{ik} \prod_{k'=1}^{k-1} (1 - v_{ik'}). \end{aligned} \tag{5}$$

While our model assumes theoretically an infinite number K of populations, given a particular finite-sized dataset, only a finite (but random) number of populations will be used to model the data, and so the posterior distribution over this number can be used to estimate the number of populations exhibited in the data.

The model is completed by specifying a prior distribution on α, α_0, r and $\theta_{kl}, k = 1, \ldots, K; l = 1, \ldots, L$. For each population k, we use independent Dirichlets for the allele frequencies at each locus. In the case of SNP data, this implies assuming independent Beta prior distribution for each locus in each subpopulation. In our simulations and our application to the ectodysplasin-A receptor (EDAR) data, we take $\theta_{kl1} \sim \text{Beta}(c\mu_l, c(1 - \mu_l))$, where μ_l denotes the prior mean for the allele frequency (assumed to be the same for all ancestral populations) and c is a concentration parameter. We choose independent Gamma prior distributions for α and α_0 for computational reasons. We specify a uniform prior on $\log r$, on a fairly large interval. Recall that d_l denotes the genetic distance between adjacent markers. If this distance is measured in Morgans, then r can be interpreted as an estimate of t, the number of generations since the admixture event. See Falush et al. (2003) for details. When the genetic distance between loci is not available, we can use as a proxy the physical distance measured in nucleotides. In this case r be interpreted as an estimate of the product of t and the recombination rate (expected number of crossovers per base pair per meiosis).

Another important issue which arises is the computational requirements for inference in a model with an infinite number of populations. In this regard, a range of recent truncation and marginalization techniques can be applied allowing for exact inference using finite computational resources. See, *e.g.* Neal (2000), Walker (2007), Papaspiliopoulos

and Roberts (2008) and Favaro et al. (2013). Before presenting our approach we first recall some preliminaries for the hierarchical Dirichlet process prior described in (3) and (4).

2.2 Hierarchical Dirichlet process

The stick-breaking prior for the overall population prevalences (4) imposes a particular ordering on the populations, in which populations with higher index have *a priori* lower prevalences. This is undesirable from a modeling perspective as the induced ordering is artificial, while from a computational perspective it is also undesirable as it introduces a label switching problem into the inference, which can slow down convergence of inference algorithms (Jasra et al., 2005; Papaspiliopoulos and Roberts, 2008). We address this issue by developing a more abstract formalism for the model based on a construction of coupled random probability measures called the hierarchical Dirichlet process and introduced in Teh et al. (2012). Specifically, let (Θ, Ω) be a measurable space. The Dirichlet process $G_0 \sim \mathrm{DP}(\alpha_0, H)$ is a random probability measure over (Θ, Ω) with the property that for any measurable partition $(A_1, \ldots, A_L)$ of Θ the random probability vector $(G_0(A_1), \ldots, G_0(A_L))$ is distributed according to a Dirichlet distribution with parameters $(\alpha_0 H(A_1), \ldots, \alpha_0 H(A_L))$ (Ferguson, 1973). The parameters of the process consist of a positive concentration parameter α_0 and a base probability measure H over (Θ, Ω). One of the noteworthy properties of the Dirichlet process is that the random probability measure G_0 is discrete almost surely, and can be written as

$$G_0 = \sum_{k=1}^{\infty} q_{0k} \delta_{\theta_k}. \tag{6}$$

The atoms $(\theta_k)_{k \geq 1}$ are independent and identically distributed according to the base probability measure H, while the atom masses are independent of the atoms, and have distribution given by the stick-breaking representation (4). Note that if all of the variants are biallelic, then θ_k is a matrix, whereas if some variants have more than two forms, then θ_k must indexed by both position and variant number, as in (2).

In the context of admixture modeling, we will suppose that each atom in G_0 corresponds to a population with allelic frequencies parameterised by the atom, while the masses correspond to the population proportions or prevalences. In other words, θ_k denotes the vector of the population specific allele frequencies for the L loci under investigation. As each individual has its own population proportions while the collection of populations are shared across individuals, we can model this using the hierarchical Dirichlet process (HDP). For each individual i, let G_i be an individual-specific atomic random probability measure. These measures are conditionally independent and identically distributed given a common base probability measure G_0, that is

$$G_i | G_0 \sim \mathrm{DP}(\alpha, G_0) \tag{7}$$

Since each atom in G_i is drawn from G_0, the collection of atoms in G_i is precisely those in G_0, while each G_i has its own specific atom masses, that is

$$G_i = \sum_{k=1}^{\infty} q_{ik} \delta_{\theta_k}, \tag{8}$$

where the masses $(q_{ik})_{k\geq 1}$ have distribution as given in (5). The HDP allows sharing of the ancestral populations among the individual distributions as the G_i place atoms at the same discrete locations determined by G_0. See Teh et al. (2012) for details.

We refer to the proposed model as HDPStructure. In summary, G_i describes the proportion of the alleles on $x_i = (x_{i1}, \ldots, x_{iL})$ coming from each of the populations, as well as the parameters of the populations. We model the sequence x_i given G_i as follows: (i) first we place segment boundaries according to an nonhomogeneous Poisson process with rate rd_l, (ii) then the alleles on each segment are generated by picking a population of origin according to G_i, and then sampling the alleles according to the population distribution. We have expressed the hierarchical prior over the population proportions in (4) and (5) as the joint distribution of atom masses in a HDP, while the atoms correspond to the population parameters. Further, while the stick-breaking representation imposes a particular ordering among the atoms, there is no ordering of atoms in the representation as random probability measures themselves. As we will see hereafter, this construction allows for an efficient MCMC algorithm for posterior simulation.

The distribution induced on the sizes of the populations of origin drawn from G_0 is equivalent Ewens' sampling formula (Pitman, 2006). This sampling formula describes the allele proportions of a locus under the assumptions of a neutral model (*i.e.*, no natural selection) and random mating (Ewens, 1972). Therefore, locally HDPStructure is an approximation of the true genetic process.

2.3 Markov Chain Monte Carlo

We describe a MCMC algorithm for posterior simulation in the HDPStructure model. The MCMC algorithm iterates between updates to the random probability measures $(G_i)_{0\leq i\leq N}$, the latent state sequences $(s_{il}, z_{il})_{1\leq i\leq N, 1\leq l<L}$, and the model parameters in turn, each update conditional upon all the other variables in the model. Updates to the random probability measures make use of the so-called Chinese restaurant franchise representation of the HDP, as well as a retrospective slice sampling technique which allows for a finite truncation to the random probability measures while retaining exactness of the procedure. Updates to the latent state sequences make use of a forward filtering-backward sampling procedure as a Metropolis-Hastings proposal distribution. Finally, updates to model parameters are straightforward one-dimensional Metropolis-Hastings updates. Detailed descriptions of these updates are included in the appendix in the supplementary material (Elliott et al., 2018). Multi-threaded MATLAB software implementing this MCMC scheme is freely available, along with code released under the BSD 2-clause open source license, at `http://BigBayes.github.io/HDPStructure`.

Updates to random probability measures

Conditioned on the model parameters and latent state sequences, the update to the random probability measures $(G_i)_{0\leq i\leq N}$ follow standard results for the hierarchical Dirichlet process in Teh et al. (2012). As noted previously, since the data is finite, the

number of populations used to model the data is finite as well. Conditioned on the latent state haplotypes $\{z_{il}\}$, suppose the number of such populations (as a function of the latent state haplotypes) is K^*. For simplicity, we may index these populations as $1, \ldots, K^*$. The random probability measures can be expressed as

$$G_0 = \sum_{k=1}^{K^*} q_{0k}\delta_{\theta_k} + w_0 G_0' \qquad\qquad G_i = \sum_{k=1}^{K^*} q_{ik}\delta_{\theta_k} + w_i G_i' \tag{9}$$

for each $i = 1, \ldots, N$, where w_i is the total mass of all other atoms in G_i, which are collected, after normalising by w_i, in a random probability measure G_i'. See Teh et al. (2012) for details.

For each $i = 1, \ldots, N$ and $k = 1, \ldots, K^*$, let n_{ik} be the number of DNA segments in sequence i assigned to population k. In the Chinese restaurant franchise representation of the HDP, we introduce a set of discrete auxiliary variables m_{ik}, taking value 0 if $n_{ik} = 0$, and values in the range $\{1, \ldots, n_{ik}\}$ when $n_{ik} > 0$. Define $n_{0k} = \sum_{i=1}^{N} m_{ik}$. Then the conditional distributions of the random probability measures given $(n_{ik}, m_{ik})_{0 \le i \le N, 1 \le k \le K^*}$ are described as follows

$$(q_{01}, \ldots, q_{0K^*}, w_0)|(n_{ik}, m_{ik}) \sim \text{Dirichlet}(n_{01}, \ldots, n_{0K^*}, \alpha_0), \tag{10}$$

$$\begin{aligned}(q_{i1}, \ldots, q_{iK^*}, w_i)|(n_{ik}, m_{ik}), (q_{01}, \ldots, q_{0K^*}), w_0 \sim \text{Dirichlet}(&\alpha q_{01} + n_{i1}, \ldots, \alpha q_{0K^*} + \\ &n_{0K^*}, \alpha w_0),\end{aligned}$$

$$\begin{aligned}G_0'|(n_{ik}, m_{ik}) &\sim \text{DP}(\alpha_0, H), \\ G_i'|(n_{ik}, m_{ik}), G_0' &\sim \text{DP}(\alpha, G_0'),\end{aligned} \tag{11}$$

where the masses form a hierarchy of finite-dimensional Dirichlet distributions while the random probability measures are independent of the masses and form a hierarchy of DPs as in the prior. We refer to Teh et al. (2012) for details.

A final point of consideration relates to the fact that the random probability measures $G_0', (G_i')$ have infinitely many atoms, so not all can be simulated explicitly with finite computational resources. We address this using a retrospective slice sampling technique to truncate the random probability measures while retaining exactness (see, *e.g.* Walker (2007), Papaspiliopoulos and Roberts (2008) and Griffin and Walker (2013)). For each individual i, we introduce an auxiliary slice variable C_i having the following conditional distribution

$$q_i^{\min} = \min_{l=1,\ldots,L} q_{iz_{il}}, \tag{12}$$

$$C_i|(n_{ik}, m_{ik}), G_0, (G_i) \sim \text{Uniform}[0, q_i^{\min}]. \tag{13}$$

The slice variables are sampled just before the latent state sequences, whose updates are described in the next subsection. Furthermore, conditioned on the slice variables, the populations whose mass fall below C_i will have zero probability to be selected when the latent state sequence for individual i is updated. Hence, as a consequence, only the finitely many atoms with mass above the minimum threshold $\min_i C_i$ need be simulated. This can be achieved by simulating G_0' and (G_i') using the hierarchical stick-breaking representation displayed in (4), (5) until the left-over mass falls below the threshold.

Updates to latent states

We use a forward-filtering backward-sampling algorithm to resample the latent state sequences one at a time. In particular, conditioned on the slice variable C_i, only populations with $q_{ik} > C_i$ will have positive probability of being selected, and so the forward-backward algorithm can be computationally tractable. However, as the slice variable depends on all latent state variables, conditioning on the slice variable introduces complex dependencies among the latent state variables which precludes an exact forward filtering algorithm. We propose instead to ignore the dependencies caused by the slice variable, and use the resulting forward-backward algorithm as a Metropolis-Hastings proposal.

Suppose there are K_i populations with proportions above the slice threshold C_i. For simplicity of exposition, we will reindex the populations such that their indices are simply $\{1, \ldots, K_i\}$. The forward-backward algorithm samples from the following proposal distribution which ignores the slice threshold C_i:

$$Q((z_{il}, s_{il})_{l=1}^{L}) \propto \mathbb{P}((z_{il}, s_{il})_{l=1}^{L}, G_i)\mathbb{P}((x_{il})_{l=1}^{L}|(z_{il}, s_{il})_{l=1}^{L}|G_i), \tag{14}$$

where the population indicators range only over $1, \ldots, K_i$. The forward filtering phase first computes the following probabilities using dynamic programming:

$$\begin{aligned} M_{bk}^{il} &= \mathbb{P}(x_{i1}, \ldots, x_{il}, s_{il} = b, z_{il} = k|(q_{ik}, \theta_k)_{k=1}^{K_i}), \\ M_{\bullet k}^{il} &= \mathbb{P}(x_{i1}, \ldots, x_{il}, z_{il} = k|(q_{ik}, \theta_k)_{k=1}^{K_i}), \\ M_{\bullet\bullet}^{il} &= \mathbb{P}(x_{i1}, \ldots, x_{il}|(q_{ik}, \theta_k)_{k=1}^{K_i}), \end{aligned} \tag{15}$$

with $b \in \{0, 1\}$ and $k \in \{1, \ldots, K_i\}$. The dynamic programme starts at $l = 1$:

$$M_{\bullet k}^{i1} = \theta_{k1x_{i1}} q_{ik}$$

and proceeds with $l = 2, \ldots, L$:

$$\begin{aligned} M_{1k}^{il} &= \theta_{klx_{il}} e^{-rd_{l-1}} M_{\bullet k}^{il-1} & M_{\bullet k}^{il} &= M_{0k}^{il} + M_{1k}^{il} \\ M_{0k}^{il} &= \theta_{klx_{il}} (1 - e^{-rd_{l-1}}) q_{ik} M_{\bullet\bullet}^{il-1} & M_{\bullet\bullet}^{il} &= \sum_{k=1}^{K_i} M_{\bullet k}^{il}. \end{aligned}$$

Recall that $\theta_{klx_{il}}$ is the probability that locus l in population k assume the observed value x_{il}. The backward phase samples from the proposal distribution, starting at $l = L$:

$$\begin{aligned} Q(z_{iL} = k) &\propto M_{\bullet k}^{iL}, \\ Q(s_{iL} = b|z_{iL} = k) &\propto M_{bk}^{iL} \end{aligned}$$

and iterates backwards, for $l = L - 1, \ldots, 1$:

$$\begin{aligned} Q(z_{il} = k|s_{il+1} = 1, z_{il+1} = k') &\propto \mathbf{1}(k = k'), \\ Q(z_{il} = k|s_{il+1} = 0, z_{il+1} = k') &\propto M_{\bullet k}^{il}, \end{aligned}$$

$$Q(s_{il} = b | z_{il} = k) \propto M_{bk}^{il},$$

where $\mathbf{1}(\cdot)$ denotes the indicator function and where $s_{i1} = 0$ by construction. In particular note that in this way we obtain a new sample for the collection of s_{il} and z_{il}. Finally, the Metropolis-Hastings acceptance probability is a simple expression which accounts for the effect of conditioning on C_i, that is

$$\min\left(1, \frac{q_i^{\text{min-cur}}}{q_i^{\text{min-prop}}}\right), \tag{16}$$

where $q_i^{\text{min-cur}}$ and $q_i^{\text{min-prop}}$ indicate the minimum population proportions (12) under the current and proposed states respectively. The forward-backward algorithm has a computational scaling of $\mathcal{O}(LK_i)$, linear in both the length of the sequence and the number of potential populations, and is the most computationally expensive part of the MCMC algorithm. It must be noted that, since the (G_i) are conditionally independent given G_0, the algorithm can be easily parallelised so as to exploit modern parallel computation technology. This parallelisation is done in the code we provide online.

2.4 Extensions

The proposed approach can be straightforwardly extended to diploid or polyploid data, by assuming that, for each individual i, the z_i along each of individual i's chromosomes form independent Markov chains satisfying (2). Other extensions of the Bayesian nonparametric admixture model can be introduced to allow correlated allele frequencies. For instance, following the approaches of Pritchard et al. (2000) and. Falush et al. (2003), it is straightforward to introduce a Bayesian nonparametric admixture model with correlated allele frequencies. Specifically, we can assume that allele frequencies in one population provide information about the allele frequencies in another population, *i.e.* frequencies in the different populations are likely to be similar (due to migration or shared ancestry). In particular this can be achieved by specifying a more complex prior structure on $\theta_{kl.}$, for example employing the correlated allele frequencies model of Falush et al. (2003), which assumes that allele frequencies at locus l in different populations are deviations from allele frequencies in a hypothetical ancestral population.

At the moment we use only genetic data to infer the admixture parameters. Often it can be useful to include in the model extra information such as physical characteristics, *e.g.*, ethnicity, of the sample individuals or geographic sampling locations as proposed in Hubisz et al. (2009). Of course these new sources of information would modify the clustering structure and would allow the proportion of individuals assigned to a particular cluster to depend on the new information. In particular this would require a specification of a spatiality dependent model on the weights of the random measures in the HDP.

From a Bayesian nonparametric perspective, we could also employ other priors such as the Pitman–Yor process introduced in Pitman and Yor (1997), and its hierarchical extension discussed in Teh and Jordan (2010). The Pitman–Yor process is a two-parameter generalization of the DP, for which a stick-breaking construction and a Chinese restaurant representation still hold. Under certain assumptions, it can be shown that the

number of clusters in a sample from a Pitman–Yor process grows much faster than for a standard DP and that the cluster sizes decay according to a power law. This property makes the Pitman–Yor process a more suitable choice in many applications. The implementation of this more flexible nonparametric prior would require more expensive computations due to the larger number of extant populations possible.

3 Illustration

We demonstrate our model on a dataset of 372 Colombian people recently genotyped on the Illumina Human610-Quadv1_B SNP array as part of a genome-wide association study (Scharf et al., 2013). South American and Central American samples are uniquely advantageous for this purpose (Ruiz-Linares et al., 2014) because of their well-documented history of extensive mixing between indigenous peoples of the Americas and people arriving from Europe and Africa. This continental admixture, which has occurred for the past 500 years (or about 20–25 generations), gives rise to haplotype blocks which are about the right length for such analysis. Ancient admixture produces very short haplotype fragments which are hard to assign ancestry with certainty, while very recent admixture allows only large haplotype blocks and there is not sufficient variation in ancestry for individuals.

The indigenous peoples of the Americas arose as a branch of the East Asian populations who were separated over 15,000 years ago and consequently isolation and genetic drift shaped their genetic landscape. This caused many SNPs to drift even more than their East Asian counterparts, eventually becoming fixed at the alternative allele. The Ectodysplasin-A receptor (EDAR) gene, located on chromosome 2, is a common example, in particular SNP rs3827760 (Mikkola, 2009), whose ancestral A allele is 100% prevalent in European and African populations (Adhikari et al., 2016b), but the alternative G allele is seen at 94% frequency in Han Chinese and 98% in indigenous peoples. The SNP, a missense mutation, has been observed to have a range of functional effects in humans (Adhikari et al., 2015, 2016b,a) and replicated in other mammals such as mice, including the characteristic straight hair shape in East Asians (Fujimoto et al., 2008; Tan et al., 2013). Our dataset does not contain rs3827760, but neighbouring SNPs in linkage disequilibrium (LD) with rs3827760 are included in the chip panel. This shows the strength of our model, as we manage to capture the ancestries even in absence of SNP rs3827760, the well-known causal and ancestry-informative SNP, by making good use of linkage disequilibrium information. Figure 1 shows the linkage disequilibrium plot (*i.e.*, the R^2 between each pair of SNPs) for the EDAR region in the Colombian samples. The LD plot was produced using the LDmatrix tool from the NIH National Cancer Institute Division of Cancer Epidemiology and Genetics[1]. Overall, EDAR signalling acts during prenatal development to specify the location, size and shape of ectodermal appendages, such as hair follicles, teeth and glands (Mikkola, 2009). Therefore, we considered EDAR to be an interesting candidate for admixture analysis as it carries information regarding ancestry due to its variation across ancestry as well as its range of functional effects.

[1] urlhttps://analysistools.nci.nih.gov/LDlink/

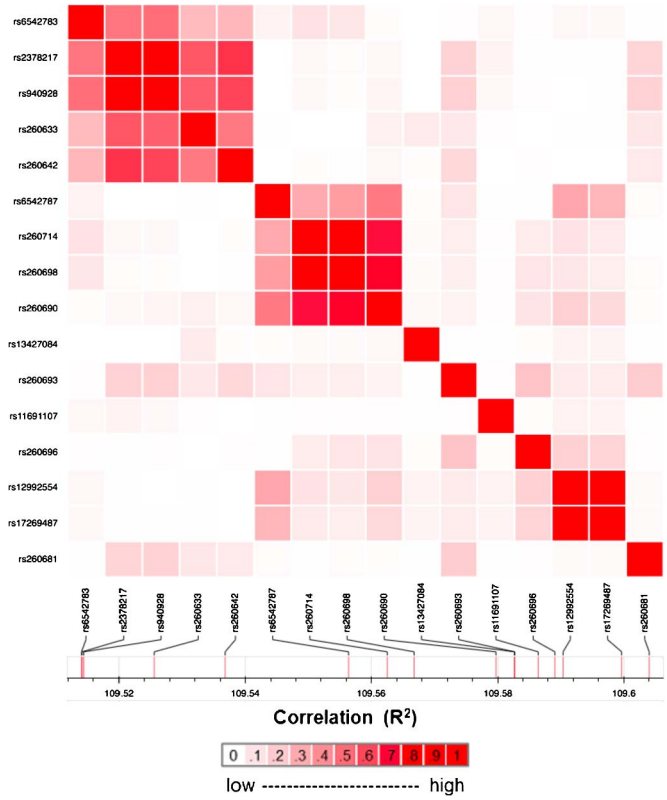

Figure 1: Linkage disequilibrium plot for the EDAR region in the Colombian samples.

Genotype information on 372 individuals for 16 SNPs in the EDAR region was available from our Illumina chip data. Genotypes were phased for conversion to haplotype format using ShapeIt2 (Delaneau et al., 2013). Data from a total of 828 individuals sampled in putative parental populations were used as reference ancestral groups. These were selected from the International Haplotype Map Project (HAPMAP), the The Centre d'Etude du Polymorphism Humain/Human Genome Diversity Project (CEPH-HDGP) cell panel (Li et al., 2008) and from published data on indigenous peoples of the Americas (Reich et al., 2012) as follows: 169 African people (from 5 populations from Sub-Saharan West Africa), 299 Europeans (from 7 West and South European populations) and 360 indigenous people (47 populations from Mexico and the Americas South of Mexico).

We ran the MCMC sampler for 50,000 iterations. We collected samples after a burn-in of 20,0000 iterations and thinned every 30 iterations. We specified the following prior distributions for the precision parameters in the HDP: $a_0 \sim \text{Gamma}(1,1), \alpha \sim \text{Gamma}(10,20)$. We centred the prior for the mean parameters of the Beta base measure of the HDP around the overall observed allele frequencies, with $c = 0.01$. The prior for $\log r$ was a Uniform on the interval $[-500, 5]$. The MCMC sampler was initialised

Correlation	European anc.	Indigenous anc.	African anc.
Cluster 1 (Europe)	0.90	-0.67	-0.36
Cluster 2 (America)	-0.75	0.92	-0.20
Cluster 3 (Africa)	-0.36	-0.19	0.76
Cluster 4 (New)	0.04	-0.24	0.28

Table 1: Correlation between ancestry proportions and cluster occurrence proportion from the Bayesian nonparametric model.

using a linkage based clustering algorithm (MATLAB's 'linkage' function), which uses agglomerative clustering to assign each subject to one of $K = 5$ initial clusters.

The posterior analysis shows evidence of four major ancestral populations in the set of 744 Colombian haplotypes (see Figure 3, right). We used the MCMC output to estimate the cluster assignment, *i.e.* population allocation, to each of the 4 major ancestral populations for each haplotype sequence and each marker. In Figure 2, we summarise the MCMC output by reporting the clustering that minimizes the posterior expectation of Binder's loss as described by Fritsch et al. (2009), who also discuss possible alternatives such as Maximum *a posteriori* clustering. The four major clusters have admixture coefficients, (*i.e.* relative proportion of occurrences of each of the clusters), 51.8%, 32.1%, 11.4% and 4.7% respectively. As we have used a reference panel, we are able to identify in the first cluster, in terms of cardinality, European-origin haplotypes in the sampled Colombian people. The second and third clusters correspond to indigenous people and African people respectively. This is also confirmed by looking at the "most frequent" haplotype in each cluster.

Figure 4 shows the population structure assigned to each of the four major ancestral populations. We verified our findings in two ways. Firstly, we calculated genetic ancestry proportions using reference genotypes as reference ancestral groups. EDAR-specific ancestry proportions for each of the 372 Colombian samples were estimated using Admixture software (Alexander et al., 2009), which provides a faster implementation of the same model that is used in STRUCTURE. We correlated these ancestry proportions to our cluster occurrence proportion (see Table 1). The correlation values are very high and support our assignment of ancestry category to the first three clusters. The average European, indigenous and African ancestry across Colombian samples are 53.6%, 30.8% and 16.6% respectively, which is also very close to our cluster proportions. However, Admixture is a supervised approach and so it cannot give us further details about the fourth, rarer cluster. To explore our results further, and also for another line of verification, we calculated genetic principal components, in which SNP genotypes for each person is recoded into 0/1/2 by an additive count of the minor allele on two chromosomes, and this SNP genotype count matrix is converted to principal components (PC) *via* the usual method. As European+indigenous continental genetic mixing is the primary source for admixture in our data, the first PC axis reflects this, being positively correlated with European samples and negatively with American samples. The second PC captures the other continental component in our samples, namely the African samples. More specifically, PC1 captures the European-indigenous axis of variation, and PC2 captures the African-European axis. In Table 2 we show correlations

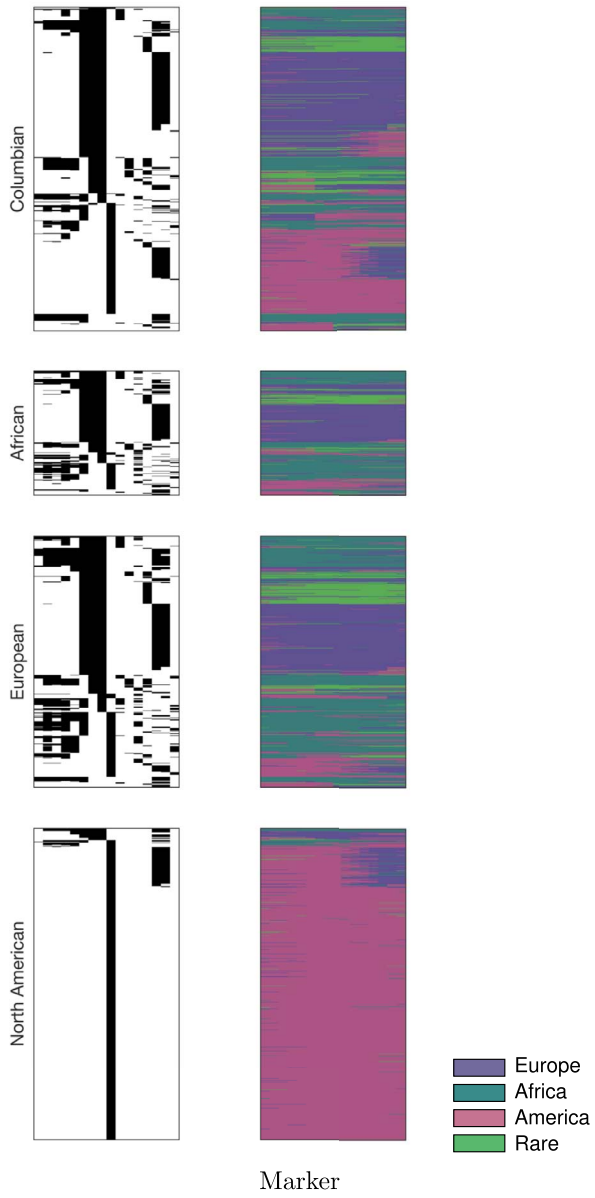

Figure 2: The left panel shows the genetic data, with black representing the minor allele for each SNP. The right panel presents summary of the posterior population assignment obtained by minimising the Binder loss function. Each row corresponds to one haplotype, with colours indicated by the legend. Individuals from each population are sorted in a lexicographical order, according to the sequence of SNPs spiraling out from the 8th SNP (*i.e.*, subjects with the minor allele at positions 8, 9, 7, 10, 6, ... are shown first).

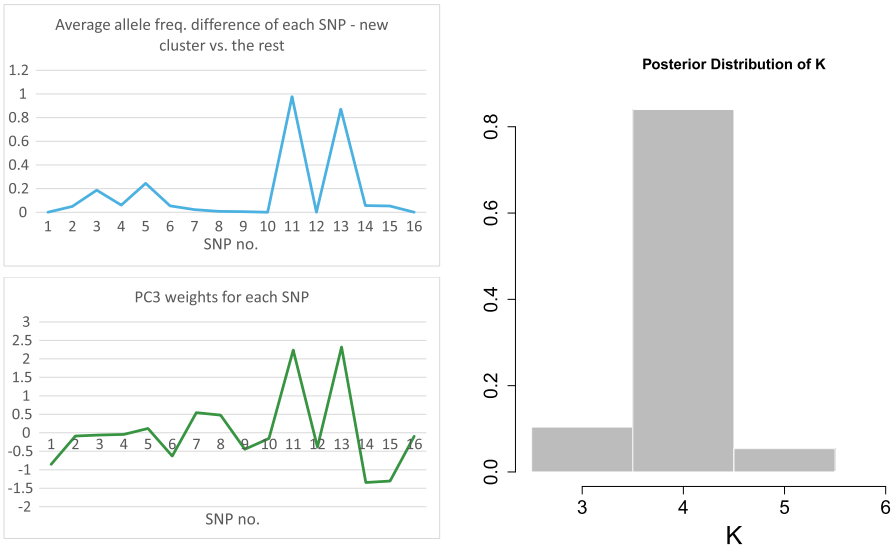

Figure 3: Left, top panel: absolute value of the difference between the average allele frequency for each SNP in all the clusters and the frequency in the rare cluster. Left, bottom panel: loadings for each SNP in the third PC. Right: EDAR data: posterior distribution of the number of clusters K.

	European anc.	Indigenous anc.	African anc.
PC1	0.81	-0.95	0.15
PC2	-0.60	-0.04	0.90
PC3	-0.09	0.00	0.13
PC4	0.15	0.03	-0.25

Table 2: Correlations between principal components and supervised ancestry values.

of PCs with supervised ancestry values. As further PCs are orthogonal to these, they do not show high correlation with any ancestry component. Consequently, the first PC shows high correlations with the first two clusters, and PC2 with the third cluster. As shown in Table 3 the third PC is highly correlated with the new cluster, which validates the signal we capture as genuine genetic component and not a statistical artefact of our method. To investigate the genetic source of the new cluster, we looked at the average allele frequency for each SNP in all the clusters, and then took the difference for the new cluster vs. all the others. Figure 3, left top panel, shows the absolute differences: we see clearly that only SNPs 11 and 13 primarily contribute to this cluster. The same is seen when we plot the weights given by PC3 onto each SNP (Figure 3 left bottom panel). These two SNPs – rs260693 and rs260696 – are rare SNPs, *i.e.* their minor allele (which has high loading in cluster 4 and PC3) is rare. For example, rs260693 has a global minor allele frequency of 3.8%, with the minor allele only primarily seen in Europe (9%) while nearly being absent in African people (1.8%) and East Asian people (0%). Their two

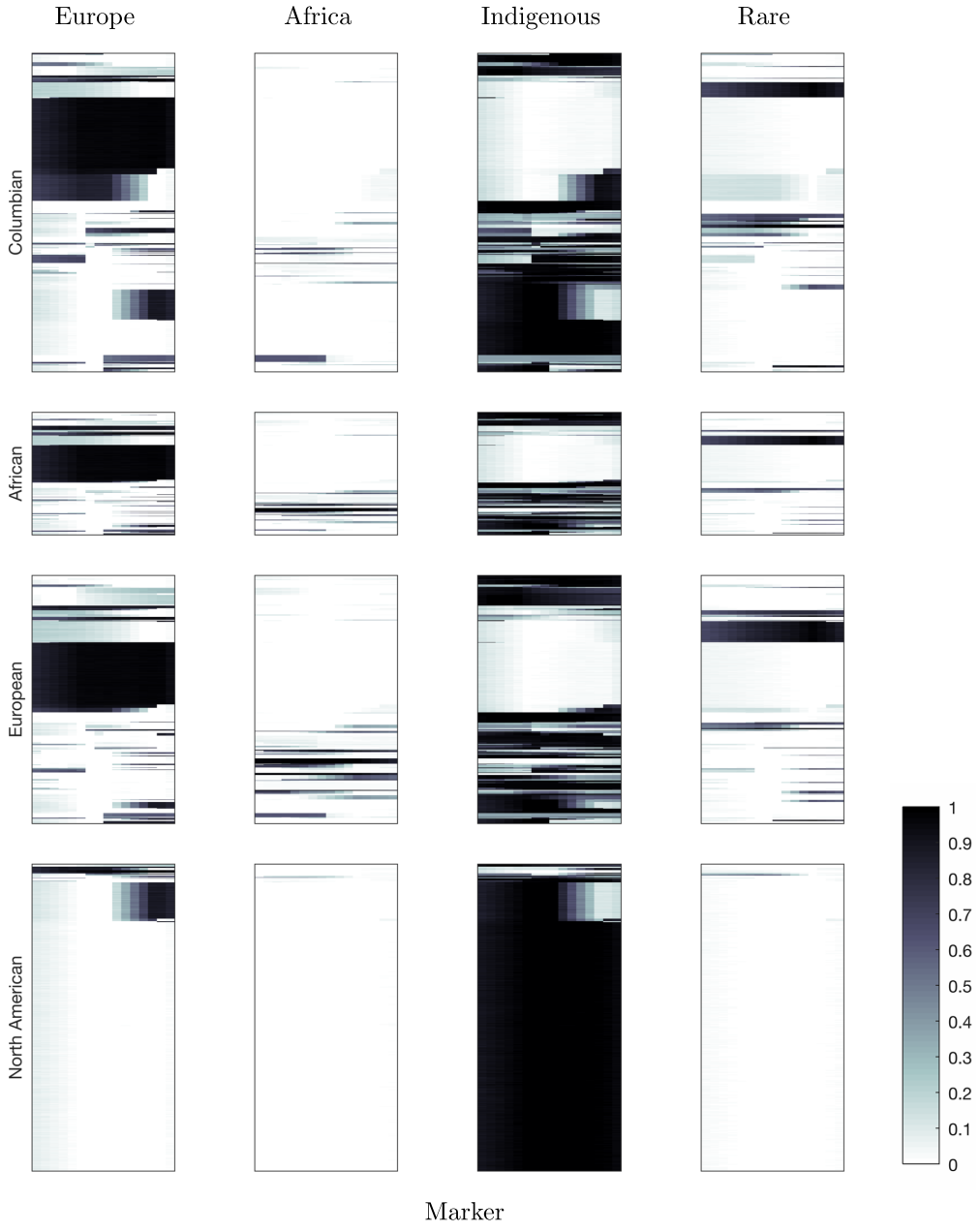

Figure 4: Each panel shows the posterior population distribution for each of the four major populations: white indicates markers for each individual not assigned to the specific population. Each row corresponds to one haplotype, with the ordering as described in Figure 3.

Correlation	PC1	PC2	PC3	PC4
Cluster 1 (Europe)	0.73	-0.58	-0.26	0.24
Cluster 2 (America)	-0.90	0.04	0.02	-0.01
Cluster 3 (Africa)	0.06	0.72	-0.07	-0.12
Cluster 4 (New)	0.24	0.31	0.71	-0.40

Table 3: Correlations between principal components and cluster assignment.

rsid	Europe	Africa	America	Rare
rs6542783	0.963	0.998	0.580	1.000
rs2378217	0.900	0.982	0.299	0.000
rs940928	0.907	0.981	0.452	0.000
rs260633	0.835	0.973	0.289	0.000
rs260642	0.877	0.968	0.539	0.001
rs6542787	0.129	0.956	0.495	1.000
rs260714	0.073	0.968	0.822	0.000
rs260698	0.063	0.970	0.624	0.000
rs260690	0.971	0.036	0.388	1.000
rs13427084	0.949	1.000	0.798	1.000
rs260693	0.943	0.996	0.985	0.001
rs11691107	0.973	1.000	0.638	1.000
rs260696	0.808	0.990	0.916	0.000
rs12992554	0.348	0.817	0.865	1.000
rs17269487	0.369	0.823	0.872	1.000
rs260681	1.000	1.000	0.738	1.000

Table 4: Posterior mean of $\theta_{kl}, l = 1, \ldots, 16$ in each of the four major populations. The colour gradient in each cell is proportional to the numerical value.

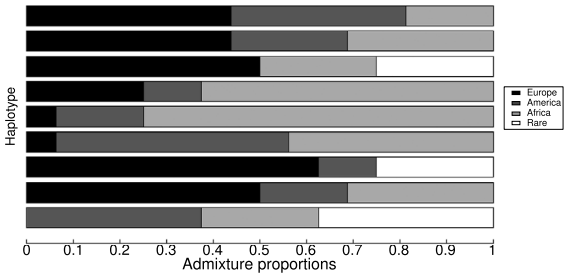

Figure 5: Posterior admixture proportions for randomly selected haplotypes.

minor alleles are in high LD, with a D' of 1 in European populations. This shows that the haplotype that contains the two minor alleles for these two SNPs is also rare, and is identified as the fourth separate cluster by our model. Table 4 reports the posterior mean of the θ_{kl} of the 16 SNPs in each of the four major populations while Figure 5 shows the posterior mean of admixture proportion q_{ik} for a few randomly selected individuals.

#threads	2	4	6	8
Runtime (seconds)	49,783	26,896	19,098	15,665

Table 5: Runtime (seconds) for HDPStructure EDAR experiment, as # threads varies, showing a speed increase of 2.2x over the range.

In Table 5, we report runtimes for the EDAR experiment described in this section. We varied the number of threads used by the HDPStructure parallelism from 2 to 8, and for each thread condition we report the median runtime over 5 restarts (with 50,000 iterations and parameter settings as described above) and found that the 8 thread condition was 3.2 times faster than the 2 thread condition, indicating a significant speedup gained through the parallelism.

3.1 Simulated experiments and comparisons

We compared the performance of HDPStructure to that of STRUCTURE2 and a baseline parametric clustering method using simulated data for which the ground truth admixture of each simulated individual is known. The data for $N = 600$ individuals and $L = 800$ SNPs were simulated according to the following generative process, which is based on a beta/Bernoulli model:

- Set global admixture proportions G_0 to the vector $(.1, .2, .3, .4)$.
- For each $i = 1, \ldots, N$, set the admixture proportions for individual i to the vector $G_i \sim \text{Dirichlet}(2.0 * G_0)$.
- Choose 20 recombination hot spots randomly by sampling the set HOT uniformly from $\{1, \ldots, L-1\}$ without replacement.
- For each individual, draw jump split points independently between each SNP according to $s_{il} \sim \text{Bernoulli}(1 - \lambda_l)$, where $l = 1, \ldots, L-1$, and $\lambda_l = 0.001$ for $l \notin \text{HOT}$, and $\lambda_l \sim \text{Uniform}(0.01, 0.5)$ for $l \in \text{HOT}$.
- For $k = 1, \ldots, K$ generate a latent haplotype for the k-th admixture component by drawing $h_{kl} \sim \text{Bernoulli}(0.25)$ independently.
- For each individual, choose the admixture component z_{il} for that individual at each SNP by drawing $z_{i1} \sim \text{Discrete}(G_i)$ for the first SNP. For the rest of the SNPs, if $s_{i,l-1} = 1$, then set z_{il} to $z_{i,l-1}$ and otherwise draw z_{il} from $\text{Discrete}(G_i)$.
- Choose a noise level η, and set the observed allele for individual i at SNP l to be $x_{il} = h_{z_{il}l}$ with probability $1 - \eta$, and $x_{il} \sim \text{Bernoulli}(0.5)$ with probability η.

This simulation procedure generates from an admixture model with linkage and hotspots (Myers et al., 2005), and it is a parametric version of HDPStructure, and also a Bayesian version of STRUCTURE2. The noise level η injects uniform uncorrelated noise, which is interpreted as *de novo* mutation, genotyping error, and unmodelled

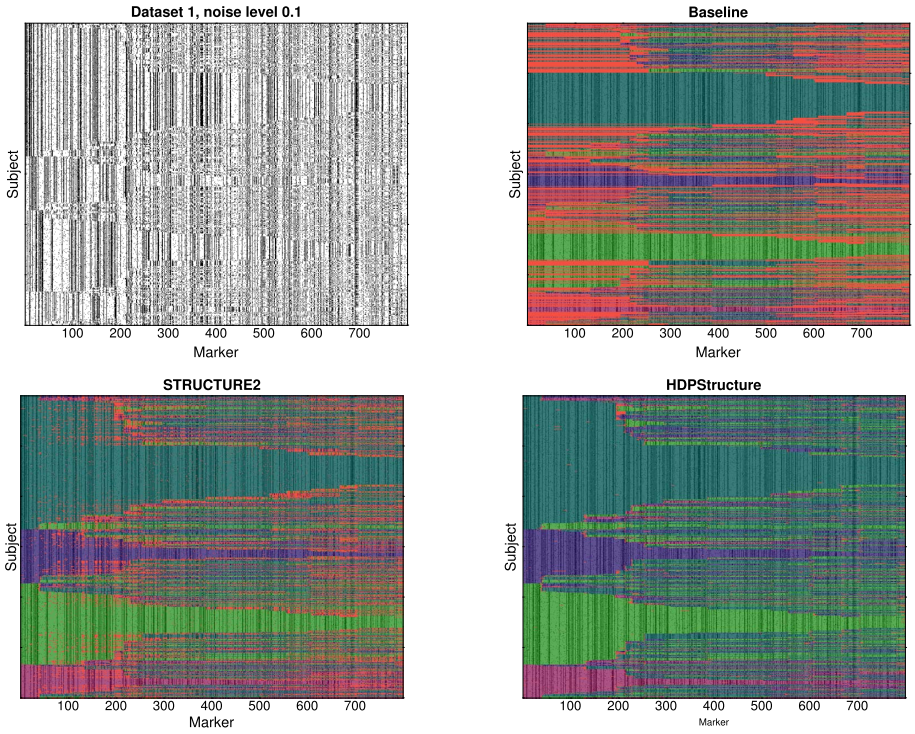

Figure 6: Results for admixture on simulated dataset, showing quality of HDPStructure's recovery. Top left: SNPs for each simulated individual. Minor alleles indicated in black. Top right, and bottom: clusterings found by baseline method, STRUCTURE2 and HDPStructure. Colour indicates index of recovered admixture components for correctly recovered entries. Red indicates mismatch between ground truth and recovered component. Order of individuals in all panes are identical and given by a lexicographical sort according to ground truth.

genetic processes. We repeat this procedure 25 times for 4 settings of the noise level: $\eta = 0.1, 0.2, 0.3, 0.4$, forming 100 simulated datasets in total. An example of one of these datasets is provided in Figure 6, top left panel. In that dataset, a hotspot ($l \in \text{HOT}$) is visable around SNP 40.

For each of these datasets, we ran the HDPStructure and STRUCTURE2 methods conditioned on the simulated SNPs x. We also formed a baseline by extracting the MATLAB linkage based clustering used to initialise HDPStructure. We then record the per-SNP admixture component assignments produced by each method. We ran STRUCTURE2 with default parameters, along with the derectives '`#define SITEBYSITE 1`' and '`#define LINKAGE 1`'. These directives indicate to STRUCTURE2 that the linkage model should be used, and that per-SNP admixture component assignments should be produced. We ran HDPStructure with settings identitcal to the EDAR experiment above, and as in the EDAR experiment, we formed the per-SNP admixture components

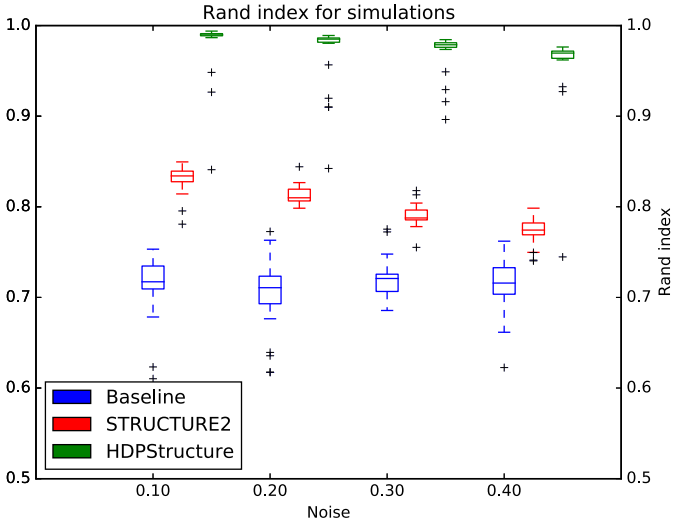

Figure 7: Rand index between baseline method, STRUCTURE2 and HDPStructure, and ground truth for simulated datasets. Improvement of HDPStructure over STRUCTURE2 is consistent over all noise conditions.

by finding the component assignments that minimised the Binder loss induced by all MCMC samples collected after burnin.

We then compared the per-SNP admixture components formed by the three methods to the ground truth. (The matrix z in the generative process is the ground truth.) We made this comparison by computing the Rand index (Rand, 1971), which is a standard measure of the agreement between two clusterings. Rand index varies from 0 to 1, with 1 meaning the clusterings are identical.

The Rand indices for HDPStructure were consistently above those of STRUCTURE2 for all noise levels. This is shown in Figure 7. In addition, the Rand indices for HDPStructure and STRUCTURE2 both dropped slightly as the noise level η was increased. The Rand index for the baseline was significantly lower than both STRUCTURE2 and HDPStructure and did not drop as the noise level increased, which is expected as the baseline did not model the splits, which produces a source of error larger than the noise. We can also confirm the quality of HDPStructure by inspecting the clustering and highlighting areas where mismatches occur. In Figure 6 right and bottom panes, a colour coded depiction of the clusterings produced by each method are provided for the first dataset in the $\eta = 0.1$ condition. It's clear from this depiction that HDPStructure recovers an almost perfect clustering, whereas STRUCTURE2 has sporatic errors, as well as systematic errors that stretch over short intervals.

We show trace plots for the MCMC runs on 6 datasets from the $\eta = 0.4$ condition in Figure 8. These plots indicate that mixing occurs quickly, within 100 iterations. We note that this fast convergence is also seen in other HMM based methods for modeling genetic variation: the default parameters for fastPHASE and BEAGLE are 25 and 10

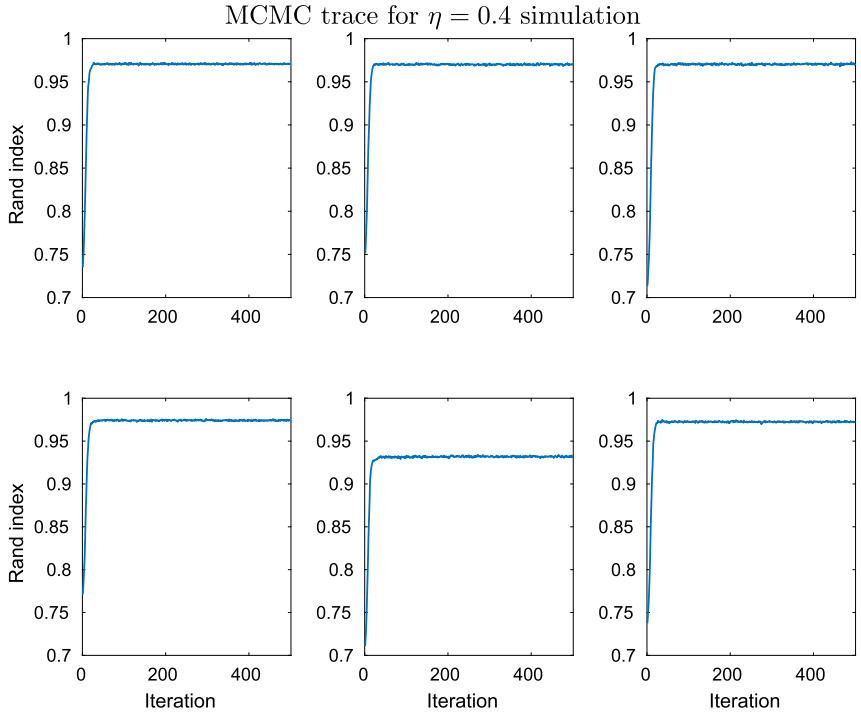

Figure 8: MCMC trace of Rand index between HDPStructure and ground truth for first 500 iterations of posterior resampling for $\eta = 0.4$ condition. Panels show trace for first 6 (of 25) datasets, and indicate convergence after ¡100 iterations.

iterations respectively (including any burnin), and IMPUTE/SHAPEIT software have a similar small number of iterations. This means that a clustering based on HDPStructure can be achieved in far fewer than the tens of thousands of iterations recommended with STRUCTURE2, although extracting posterior estimates may require these extra iterations.

The median runtime for all of the runs of HDPStructure on the simulated data was 14,037 seconds. We used one thread for each of these HDPStructure runs. The median runtime for the runs of STRUCTURE2 was slightly more, at 14,420 seconds. The number of iterations used was matched for all the HDPStructure and STRUCTURE2 runs. This means that the runtime of HDPStructure is comperable to that of STRUCTURE2, although the parallelism described in Table 5 could improve the HDPStructure runtimes further.

4 Discussion and concluding remarks

We have presented the Bayesian nonparametric counterpart of the linkage model of Falush et al. (2003) to infer genetic admixtures. The model allows for both the number

of ancestral population and the assignment vector to be random, avoiding the use of model selection criteria. The model can be applied to commonly used genetic markers and does not rely on specific assumption on the mutation model. We incorporate dependence between markers due to correlation of ancestry by specifying an inhomogeneous Poisson process on the DNA sequence. Each population is modelled using a simple and independent allele-frequency profile. We have developed an MCMC algorithm which allows us to perform posterior inference on the number of ancestral populations, the population of origin of chromosomal regions, the proportion of an individual's genome coming from each population, and the admixture proportions in the population and the allele frequencies in ancestral populations. We have demonstrated the model on real data from the EDAR gene. The model has been able to highlight the existence of a rare European haplotype. We have highlighted possible extensions to our method. An interesting direction for future development is to relax the assumption of independent allele-frequency profile in each population by incorporating ideas from Sohn et al. (2012) and model each population as a hidden Markov model over a set of founder haplotypes.

We also showed improved performance of HDPStructure over (Falush et al., 2003) in the recovery of admixture components for simulated data. This improvement could be due to the fact that the HDPStructure prior is an approximation of the true genetic process and because large populations have large prior support under the HDPStructure model, whereas in STRUCTURE2 the population sizes are optimized as parameters and have no prior.

In this article we have devoted considerable attention to inferring K and shown how Bayesian nonparamertic methods automatically provide posterior inference on the number of ancestral populations. Nevertheless, we must be careful when interpreting K. The nonparametric setup will generally yield sensible clustering but clusters will not necessarily correspond to "real" populations. This consideration also holds for other model-based structure algorithms (Pritchard et al., 2000).

Supplementary Material

Modeling population structure under hierarchical Dirichlet processes: Appendix (DOI: 10.1214/17-BA1093SUPP; .pdf).

References

Adhikari, K., Fontanil, T., Cal, S., Mendoza-Revilla, J., Fuentes-Guajardo, M., Chacón-Duque, J.-C., Al-Saadi, F., Johansson, J. A., Quinto-Sanchez, M., Acuña-Alonzo, V., et al. (2016a). "A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features." *Nature Communications*, 7. 324

Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Chacón-Duque, J. C., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Lozano, R. B., Pérez, G. M., et al. (2016b). "A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation." *Nature Communications*, 7. 324

Adhikari, K., Reales, G., Smith, A. J., Konka, E., Palmen, J., Quinto-Sanchez, M., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Fuentes, M., et al. (2015). "A genome-wide association study identifies multiple loci for variation in human ear morphology." *Nature Communications*, 6. 324

Alexander, D. H., Novembre, J., and Lange, K. (2009). "Fast model-based estimation of ancestry in unrelated individuals." *Genome Research*, 19(9): 1655–1664. 314, 326

Anderson, E. and Thompson, E. (2002). "A model-based method for identifying species hybrids using multilocus genetic data." *Genetics*, 160(3): 1217–1229. 317

Anderson, E. C. (2001). "Monte Carlo methods for inference in population genetic models." Ph.D. thesis, University of Washington. 317

Balding, D. J. and Nichols, R. A. (1995). "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity." In *Human Identification: The Use of DNA Markers*, 3–12. Springer. 317

Corander, J., Waldmann, P., and Sillanpää, M. J. (2003). "Bayesian analysis of genetic differentiation between populations." *Genetics*, 163(1): 367–374. 315

Dawson, K. J. and Belkhir, K. (2001). "A Bayesian approach to the identification of panmictic populations and the assignment of individuals." *Genetical Research*, 78(01): 59–77. 315

Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). "Improved whole-chromosome phasing for disease and population genetic studies." *Nature Methods*, 10(1): 5–6. 325

Elliott, L. T. and Teh, Y. W. (2012). "Scalable imputation of genetic data with a discrete fragmentation-coagulation process." In *Advances in Neural Information Processing Systems*, volume 24. 316

Elliott, L. T. and Teh, Y. W. (2016). "A nonparametric HMM for genetic imputation and coalescent inference." *Electronic Journal of Statistics*, 10(2): 3425–3451. 316

Elliott, L. T., De Iorio, M., Favaro, S., Adhikari, K., and Teh, Y. W. (2018). "Modeling population structure under hierarchical Dirichlet processes: Appendix." *Bayesian Analysis*. 320

Evanno, G., Regnaut, S., and Goudet, J. (2005). "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." *Molecular Ecology*, 14(8): 2611–2620. 315

Ewens, W. J. (1972). "The sampling theory of selectively neutral alleles." *Theoretical Population Biology*, 3(1): 87–112. 320

Falush, D., Stephens, M., and Pritchard, J. K. (2003). "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies." *Genetics*, 164(4): 1567–1587. 315, 316, 317, 318, 323, 334, 335

Favaro, S., Teh, Y. W., et al. (2013). "MCMC for normalized random measure mixture models." *Statistical Science*, 28(3): 335–359. 319

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. 319

Field, D., Ayre, D., Whelan, R., and Young, A. (2011). "Patterns of hybridization and asymmetrical gene flow in hybrid zones of the rare Eucalyptus aggregata and common E. rubida." *Heredity*, 106(5): 841–853. 314

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2012). "Joint modeling of multiple related time series via the beta process." http://arxiv.org/abs/1111.4226. 316

Fritsch, A., Ickstadt, K., et al. (2009). "Improved criteria for clustering based on the posterior similarity matrix." *Bayesian Analysis*, 4(2): 367–391. 326

Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., Mustofa, M. S., Samakkarn, U., Settheetham-Ishida, W., Ishida, T., et al. (2008). "A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness." *Human Molecular Genetics*, 17(6): 835–843. 324

Griffin, J. E. and Walker, S. G. (2013). "On adaptive Metropolis–Hastings methods." *Statistics and Computing*, 23(1): 123–134. 321

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press. 315

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). "Inferring weak population structure with the assistance of sample group information." *Molecular Ecology Resources*, 9(5): 1322–1332. 323

Huelsenbeck, J. P. and Andolfatto, P. (2007). "Inference of population structure under a Dirichlet process model." *Genetics*, 175(4): 1787–1802. 315

Jasra, A., Holmes, C., and Stephens, D. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling." *Statistical Science*, 20(1): 50–67. 319

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., et al. (2008). "Worldwide human relationships inferred from genome-wide patterns of variation." *Science*, 319(5866): 1100–1104. 325

Mikkola, M. L. (2009). "Molecular aspects of hypohidrotic ectodermal dysplasia." *American Journal of Medical Genetics Part A*, 149(9): 2031–2036. 324

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). "A fine-scale map of recombination rates and hotspots across the human genome." *Science*, 310(321): 321–324. 331

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. 318

Neal, R. M. (2003). "Density modeling and clustering using Dirichlet diffusion trees." *Bayesian Statistics*, 7: 619–29. 316

Novembre, J. and Stephens, M. (2008). "Interpreting principal component analyses of spatial population genetic variation." *Nature Genetics*, 40(5): 646–649. 314

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186. 318, 319, 321

Parker, H. G., Kim, L. V., Sutter, N. B., Carlson, S., Lorentzen, T. D., Malek, T. B., Johnson, G. S., DeFrance, H. B., Ostrander, E. A., and Kruglyak, L. (2004). "Genetic structure of the purebred domestic dog." *Science*, 304(5674): 1160–1164. 314

Patterson, N., Price, A. L., and Reich, D. (2006). "Population structure and eigenanalysis." *PLoS Genetics*, 2(12). 314

Pella, J. and Masuda, M. (2006). "The gibbs and split merge sampler for population mixture analysis from genetic data with incomplete baselines." *Canadian Journal of Fisheries and Aquatic Sciences*, 63(3): 576–596. 315

Pitman, J. (2006). *Combinatorial stochastic processes*. Springer-Verlag. 320

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25(2): 855–900. 323

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). "New approaches to population stratification in genome-wide association studies." *Nature Reviews Genetics*, 11(7): 459–463. 314

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). "Inference of population structure using multilocus genotype data." *Genetics*, 155(2): 945–959. 314, 315, 317, 323, 335

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66(336): 846–850. 333

Rannala, B. and Mountain, J. L. (1997). "Detecting immigration by using multilocus genotypes." *Proceedings of the National Academy of Sciences*, 94(17): 9197–9201. 317

Ray, A. and Quader, S. (2014). "Genetic diversity and population structure of Lantana camara in India indicates multiple introductions and gene flow." *Plant Biology*, 16(3): 651–658. 314

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M. V., Rojas, W., Duque, C., Mesa, N., et al. (2012). "Reconstructing native American population history." *Nature*, 488(7411): 370–374. 325

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). "Reconstructing Indian population history." *Nature*, 461(7263): 489–494. 314

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivo-

tovsky, L. A., and Feldman, M. W. (2002). "Genetic structure of human populations." *Science*, 298(5602): 2381–2385. 314

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., et al. (2014). "Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals." *PLoS Genetics*, 10(9). 324

Scharf, J. M., Yu, D., Mathews, C. A., Neale, B. M., Stewart, S. E., Fagerness, J. A., Evans, P., Gamazon, E., Edlund, C. K., Service, S., et al. (2013). "Genome-wide association study of Tourette's syndrome." *Molecular Psychiatry*, 18(6): 721–728. 324

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4(2): 639–650. 318

Sohn, K.-A., Ghahramani, Z., and Xing, E. P. (2012). "Robust estimation of local genetic ancestry in admixed populations using a nonparametric Bayesian approach." *Genetics*, 191(4): 1295–1308. 335

Sohn, K.-A. and Xing, E. P. (2007). "Hidden Markov Dirichlet process: modeling genetic inference in open ancestral space." *Bayesian Analysis*, 2(3): 501–527. 316

Tan, J., Yang, Y., Tang, K., Sabeti, P. C., Jin, L., and Wang, S. (2013). "The adaptive variant EDARV370A is associated with straight hair in East Asians." *Human Genetics*, 132(10): 1187–1191. 324

Tang, H., Peng, J., Wang, P., and Risch, N. J. (2005). "Estimation of individual admixture: analytical and study design considerations." *Genetic Epidemiology*, 28(4): 289–301. 314

Teh, Y. W. and Jordan, M. I. (2010). *Hierarchical Bayesian nonparametric models with applications*. Cambridge University Press. 323

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. 315, 319, 320, 321

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics—Simulation and Computation*, 36(1): 45–54. 318, 321

Wasser, S. K., Mailand, C., Booth, R., Mutayoba, B., Kisamo, E., Clark, B., and Stephens, M. (2007). "Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban." *Proceedings of the National Academy of Sciences*, 104(10): 4228–4233. 314

Zhu, X., Tang, H., and Risch, N. (2008). "Admixture mapping and the role of population structure for localizing disease genes." *Advances in Genetics*, 60: 547–569. 314